UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper
Economic Series 16-06
April 2016
ISSN 2340-5031

Departamento de Economía
Universidad Carlos III de Madrid
C/Madrid, 126 28903 Getafe (Spain)
Fax (34-91) 6249875

The causal effects of an intensified curriculum on cognitive skills: Evidence from a natural experiment

Vincenzo Andrietti[11]

Abstract

This paper exploits a unique universal educational policy - implemented in most German states between 2001 and 2008 - that compressed the academic-track high school curriculum into a (one-year) shorter time span, thereby increasing time of instruction and share of curriculum taught per grade. Using 2000-2012 PISA data and a quasi-experimental approach, I estimate the impacts of this intensified curriculum on cognitive skills. I find robust evidence that the reform improved, on average, the reading, mathematical, and scientific literacy skills acquired by academic-track ninth-graders upon treatment. However, I also provide evidence that the reform widened the gap in student performance with respect to parental migration background and student ability. Finally, although the reform did not affect, on average, high school grade retention, I find that the latter increased for students with parental migration background. Taken together, these findings suggest that moving to a compressed high-school curriculum did not compromise and benefited, on average, students' cognitive skills. However, they also raise equity concerns that policy-makers should be aware of.

***Keywords**:* G8 reform, Intensified curriculum, Learning intensity, Instruction time, Cognitive skills, Academic-track high school, Grade retention, Remedial education

***JEL Classification**:* I21, I28, D04

***Acknowledgments**:* This paper is an updated version of the paper *The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment*, first published in June 2015 as UC3M Working paper economic series 15-06. I wish to thank Xuejuan Su, Jan Stuhler, Vincent Hildebrand, and Julio Caceres Delpiano as well as seminar participants at IQB Berlin, Universidad Carlos III de Madrid, University of Toronto, CEA conference 2015, SOLE-EALE world conference 2015, and EEA conference 2015 for helpful comments and suggestions.

[1] Università D'Annunzio, Dipartimento di Scienze Filosofiche, Pedagogiche ed Economico-Quantitative (DiSFPEQ), Viale Pindaro 42, 65127 Pescara, Italy. E-mail: vincenzo.andrietti@unich.it.

1 Introduction

High school duration and curricula design are key features of the school system since they shape the workload distribution across grades – i.e., the amount of instruction time and the share of the overall curriculum taught per grade – and the learning intensity that students have to cope with, and might affect both the level and the distribution of students' cognitive skills. This is an important matter, given the impact that these skills have been shown to have both at the micro – e.g., on individual earnings and educational attainment (Heckman et al., 2006) – and at the macro level – e.g., on economic growth (Hanushek and Wössmann, 2008, 2011, 2012, 2015). Nonetheless, despite the recently increased interest in the role of instruction time as an educational input (Hanushek, 2015), little attention has been paid so far to settings where the increase in time of instruction pairs with an increase in the intensity of learning – i.e., with a higher per-week and per-grade share of the overall curriculum –, and might therefore create an increased burden on students.

My study addresses this question by exploiting a unique universal educational policy that reduced high school duration in most German states from nine (G9) to eight (G8) years, and compressed the instructional time and curriculum distribution into a (one-year) shorter time span. Under the G8 regime, instruction time increased on average by about 2.5 hours per week (or about 8.5 percent) over grades five to nine.[1] As a consequence, compared to their G9 counterparts, G8 ninth graders received, on average, about 95 additional hours of instruction per grade (2.5 hours per week over 38 school weeks) over grades five to nine. The additional instructional time was used to teach learning content that was previously taught in higher grades. G8 students were therefore exposed to an intensified curriculum – i.e., they had to learn a higher per-week and per-grade share of the overall curriculum – and experienced an increase in the intensity of learning.

Using pooled cross-sectional data from the German extensions of the Programme for International Student Assessment (PISA) and a quasi-experimental approach, I investi-

[1]The increase was higher in grades seven to nine (about 3.1 hours, or 10 percent of the average baseline), and lower in grades five and six (about 1.5 hours, or 5 percent). See Section 2.

gate the impacts of this intensified curriculum on the literacy skills of academic-track ninth-graders in reading, mathematics, and science. I find positive and significant effects in all the domains tested: Reading and mathematics scores increased, on average, by 0.073 standard deviations; science scores increased by 0.087 standard deviations. Overall, these results are very robust to a variety of robustness checks with respect to possible threats to the internal validity of my quasi-experimental design.

Besides estimating the effects of a G8 treatment variable that discretely switches off and on, I examine the effects of different measures of treatment intensity, such as the duration of treatment and the average year-week hours of instruction allocated across grades five to nine. The estimation results corroborate my main findings: Depending on the domain, each additional year of exposure to treatment led to a 0.013-0.016 standard deviations increase in test scores (e.g., the modal treatment duration (five years) increased test scores by 0.065-0.08 standard deviations); or, 2.5 additional year-week hours of instruction delivered from grade five to grade nine improved test scores by 0.058-0.08 standard deviations.

Moreover, to shed further light on the effects of the reform, I estimate additional specifications that explore possible heterogeneous policy effects. I find that the reform effects are driven by girls in reading, and by students with no parental migration background and high achieving students in all the domains tested.

Finally, I explore further margins that might indicate potential unintended effects of the reform. I find no evidence of a significant average effect of the reform on high school grade retention. However, I do find that the latter increased significantly for students with parental migration background. In contrast, the reform reduced the probability of receiving remedial education in math, or in any subject.

This study contributes to three strands of literature. First, it adds to the literature analyzing the role of instruction time on student achievement. Although recent research generally supports the notion that additional instruction time increases student achievement, difficulties in isolating an exogenous source of variation raise concerns about the

strength of much of the evidence.[2] Lavy (2015) and Rivkin and Schiman (2015) address this issue by exploiting within-student variation in subject-specific instruction time delivered during the grade (ninth or tenth) attended by 15-year-old students assessed in PISA 2006 and PISA 2009, respectively. In contrast, I overcome these difficulties through a research design that exploits the increase in instruction time allotted across the early grades (five to nine) of the academic-track high school curriculum under the G8 regime.

Second, I focus on a type of quasi-experimental variation in quantity of instruction time that differs from the one typically considered in the literature. Earlier studies exploit the exogenous variation in instruction time offered by policies that lengthen the school day (Bellei, 2009; Lavy, 2012; Kraft, 2015) or the school year (Parinduri, 2014), shift state-mandated school start and/or test dates (Sims, 2008; Fitzpatrick et al., 2011; Hansen, 2011; Agüero and Beleche, 2013; Aucejo and Romano, 2014; Carlsson et al., 2015), or reallocate instruction time into a specific subject (Allensworth et al., 2009; Nomi and Allensworth, 2009, 2013; Cortes and Goodman, 2014; Taylor, 2014; Cortes et al., 2015; Dougherty, 2015) or into a shorter school week (Anderson and Walker, 2015). These policies are typically short-lived and, most importantly, do not alter the share of the overall curriculum covered in school week or in a school year, i.e., the intensity of learning.[3] Consistent with the idea that students might benefit from additional time of instruction used by teachers to cover the same curricular content in more depth – i.e., with more opportunities for practice and review – or to support slow learners, these studies generally find positive (albeit sometimes small) and significant instruction-time effects. In contrast, I exploit a reform that induced a large and lasting increase in instruction time: Most importantly, the additional time of instruction was used to teach shares of the curriculum that were previously taught in higher grades, thereby increasing the intensity

[2]For example, a few correlation studies exploit between-country variation in instruction time, finding small positive effects (Wössman, 2003) or no effects (Lee and Barro, 2001). Dobbie and Fryer (2013) find that New York City's charter schools that add 25 percent or more instruction time have annual gains that are 0.05 standard deviations higher in math. They caution, though, against a causal interpretation of their findings, due to the lack of exogenous variation in instructionl time.

[3]Similarly, policies that reduce instruction time by shortening high school duration and the corresponding curriculum (Morin, 2013; Krashinsky, 2014), or natural events that reduce instruction time by producing unscheduled school closings (Marcotte, 2007; Marcotte and Hemelt, 2008; Goodman, 2014), leave unaltered the intensity of learning.

of learning.[4] Furthermore, while the G8 reform is a universal policy (targeting the population of academic-track high school students) that alters time (and timing) of instruction without directly affecting other school inputs, in some of the previously mentioned studies the increase in instruction time is either part of a remedial intervention that directly alters the peer group composition (Allensworth et al., 2009; Nomi and Allensworth, 2009, 2013; Cortes and Goodman, 2014; Taylor, 2014; Cortes et al., 2015; Dougherty, 2015), or part of a bundle of policies that directly affect other school inputs (Bellei, 2009; Lavy, 2012).

Finally, my study contributes to the existing G8 literature along several dimensions. First, by exploiting an additional comparison group (i.e., middle-track students), I complement the difference-in-differences research design adopted in earlier studies (Dahmann and Anger, 2014; Dahmann, 2015; Dörsam and Lauber, 2015; Huebener and Marcus, 2015; Meyer et al., 2015) – which focus on outcomes different from the ones analyzed in this study – with a difference-in-difference-in-differences approach. This alternative research design improves the strength of my identification strategy, buttressing a causal interpretation of my main difference-in-differences results. Second, my analysis is based on large samples – covering an extended time period (2000-2012) in which the G8 reform was implemented in most German states – and includes grade repeaters. In contrast, most G8 studies (see, among others, Dahmann and Anger, 2014; Büttner and Thomsen, 2015; Dahmann, 2015) are based on small samples and exclude grade repeaters. While the samples used in these studies might exhibit compositional differences between treatment and control units caused by the reform itself – e.g., in high school grade retention –, I explicitly address this possibility in my empirical analysis. Finally, and most importantly, under my research design, I am essentially considering the reform impact on the achievements of students that, by the end of grade nine post-reform, have received a considerably

[4] In a seminal study, Pischke (2007) analyzes the effects of a reform that introduced an earlier start of the academic year in 1960s Germany by shortening two contiguous academic years (1966-1967). Similar to the G8 reform, the reform dramatically changed the amount of instruction time for some students in school at the time without directly affecting the curriculum, thereby increasing the learning intensity (i.e., the same curriculum had to be covered in a shorter time for the grades affected). This change increased grade repetition in primary school and lowered enrollment in academic-track high school, but it had no adverse consequences on earnings and employment outcomes of the affected cohorts.

higher amount of instruction time,[5] and covered a higher share of the overall curriculum, than students that have completed grade nine pre-reform.[6] As a consequence, this study is not about the overall effect of the G8 reform (i.e., higher intensity and shorter duration), but only focuses on its higher intensity aspect.[7] Although at first blush the focus on short-term student outcomes might be considered a limitation of this study, the importance of assessing the impact of educational reforms on standardized measures of cognitive skills – like PISA scores – is advocated by a large literature (see, among others, Hanushek and Wössmann, 2008, 2011, 2012, 2015) that posits a positive causal link between these and longer term outcomes such as individual earnings, the distribution of income, and economic growth.

While I provide robust evidence that academic-track ninth graders treated by the reform benefited, on average, from the additional instruction time allotted across grades five to nine, the benefits appear to be small, when compared to the large and lasting increase in instruction time they were exposed to, or to the typical average gains in PISA scores from an additional lower secondary school year in Germany. Besides the intuitive argument that time of instruction might have diminishing returns (Rivkin and Schiman, 2015), the higher intensity of learning might be important in explaining this finding: Students with lower initial skills might not be able to deal with the increased pace of learning (Cuhna and Heckman, 2007); or, it might be the case that students experience difficulties in absorbing the additional knowledge that they are exposed to in earlier grades (Clotfelter et al., 2015; Dougherty et al., 2015). Also, while students may realize these benefits each year, prior research on other educational interventions suggests that

[5] About 475 additional hours allocated across grades five to nine (2.5 hours per week $\times$ 38 school weeks $\times$ 5 grades), or 41 percent of the instruction time (1,150 hours) allocated on average to each of these grades pre-reform.

[6] This comparison is particularly meaningful, given that PISA tests focus on general cognitive skills, rather than mastery of specific curricular content (OECD, 2003; Hanushek and Wössmann, 2008).

[7] A number of studies (Thile et al., 2014; Büttner and Thomsen, 2015; Dahmann, 2015; Meyer and Thomsen, 2016) compare instead graduation or post-graduation outcomes of pre- and post-reform students from a single state-specific double cohort, providing different accounts of the overall reform effects. Thile et al. (2014) and Dahmann (2015) find no significant reform effects on non-cognitive and cognitive skills, respectively. In contrast, Meyer and Thomsen (2016) find a significant delay in university enrollment among female students, and Büttner and Thomsen (2015) find a significant negative effect on final achievement in mathematics. My study also adds to this literature by narrowing its focus to the effects of the increased learning intensity introduced by the G8 reform on shorter-term student outcomes.

the impacts on test scores may fade out significantly over time (see, among others, Jacob et al., 2010).

Furthermore, if the objective of school reform is to increase achievement for all students and simultaneously close the achievement gap between lower and higher achieving students, the widened achievement gaps that I find with respect to parental migration background and student ability raise equity concerns, suggesting that an analysis of the distributional effects associated with the reform may offer important policy insights. While the main focus of this paper is the "average" treatment effect of the G8 reform, I examine this issue in a separate paper. In Andrietti and Su (2016), we propose a theoretical model of the match between education curriculum and student initial preparation, and analyze the distributional impact of a change in the education curriculum on student achievement. Taking advantage of the quasi-experimental nature of the intensified curriculum introduced by the G8 reform, we then test the model predictions estimating conditional and unconditional quantile treatment effects in a non-linear DiD setting. We find evidence of heterogeneous reform effects broadly consistent with our theory: While the G8 reform improves student test scores on average, such a benefit is much more pronounced for well-prepared students; in contrast, less prepared students do not benefit from the reform.[8]

Taken together, these findings suggest that moving to a compressed high-school curriculum did not compromise and, on average, improved students' cognitive skills. However, they also raise equity concerns that policy-makers should be aware of in designing or reforming high school curricula.

Section 2 provides background on the G8 reform. Section 3 illustrates the empirical strategy. Section 4 describes the data. The main results are presented in Section 5. Section 6 probes the robustness of the findings. Section 7 analyzes possible unintended effects of the reform on further margins. Section 8 concludes.

[8]In a recent working paper, Huebener et al. (2016) use PISA 2000-2012 data to answer similar questions I have examined in Andrietti (2015) and in Andrietti and Su (2016), and find similar results.

2 The G8 Reform

Educational policy in the Federal Republic of Germany is under the responsibility of the sixteen federal states. In general, children enroll in primary school at the age of six. They continue on to secondary school after four years.[9] Students are then tracked into three basic types of secondary school, each offering a single educational track geared toward the attainment of a specific school-leaving certificate.[10] The basic-track school (*Hauptschule*) and the middle-track school (*Realschule*) provide schooling through grade nine or ten, grade nine being the minimum attendance requirement in Germany. The highest level of secondary school is academic-track high school (*Gymnasium*), referred to as academic-track because only its successful completion leads to university entrance qualification (*Abitur*).

Up to 2001, the academic-track high school lasted nine years in almost all federal states, resulting in a total of thirteen years of schooling to graduate from high school and qualify for university entrance.[11] However, following a heated debate, and guided by the desire to speed up graduation and increase labor market participation of high school students, starting in 2001 most German states reduced the length of the academic-track by one year, as illustrated in Figure 1. Figure 1A displays the timing of the reform introduction, as well as the grades initially treated. Although in most states the reform affected only students entering the academic-track – i.e., fifth-graders –, some states (Saxony-Anhalt in 2003 and Bavaria, Mecklenburg-Vorpommern, and Lower Saxony in 2004) extended its applicability to students that had entered high school in previous

[9] Exceptions are the states of Berlin, Brandenburg, and (since 2007) Mecklenburg-Vorpommern, where the transition to secondary school (tracking) takes place at the start of grade seven, as opposed to grade five. In contrast, tracking in grade seven was abolished (since 2004) in Bremen and (since 2003) in Lower-Saxony (KMK, 1997-2014).

[10] Some states also have comprehensive schools (*Gesamtschulen*), which combine the three basic secondary school types in one organizational unit offering multiple educational tracks. In addition, some states offer types of school that bring the lower tracks – i.e., basic- and middle-track – under one educational and organizational umbrella. These schools – classified for statistical purposes as *schularten mit mehrerhen bildungsgängen* (schools with multiple educational tracks) – take usually state-specific names (Lohmar and Eckhardt, 2010).

[11] Whereas since the Second World War the overall length of Gymnasium in the West German states has been thirteen years, it was set at twelve years in the former East German states. Following reunification, the former East German states – with the exception of Saxony and Thuringia – adapted to West German standards, increasing the overall schooling length to thirteen years (Kühn et al., 2013).

years and currently attending upper grades (up to grade nine). Figure 1B indicates the expected graduation year of the first treated cohort in each state. The latter is considered part of a *double graduating cohort* because it is expected to graduate at the same time as the last G9 cohort. Figure 2 adds a spatial dimension, suggesting that the timing of the G8 reform implementation did not follow a geographical pattern, possibly related to region-specific economic and/or school conditions.

Under G8, the overall curriculum and the instruction time – 265 year-week hours[12] – required to cover it under the G9 regime were left unaltered, but had to be reallocated across fewer grades. As a consequence, the number of year-week hours of instruction and the corresponding share of curriculum covered per grade increased. The actual allocation policy was left up to the federal states. Figure 3 – based on state-specific official historical timetables provided by the *Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany* (KMK, 1997-2014) – compares the average year-week hours of instruction allotted to grades five to twelve under the new (G8) and the old (G9) regime. It reveals that, under G8, instructional time allocated across grades seven to nine increased on average by 3.1 hours per week (or about 10 percent of the average baseline). The increase was lower (1.5 hour per week, or about 5 percent of the baseline) in grades five and six. Thus, by the end of grade nine, G8 students were taught on average about 12.5 additional year-week hours of instruction (or, when multiplied by 38 school-weeks, about 475 hours more then their G9 counterparts).[13] These additional hours were used to teach new learning content, covering shares of the curriculum previously taught in higher grades.

[12]Year-week hours are the hours of instruction per week allocated to each academic-track high school year (grade) that are summed up over all years until graduation. A sum across grades of 265 year-week hours of instruction is considered as the minimum graduation requirement.

[13]This visual evidence is corroborated by the OLS baseline regression estimates presented in Table A.1.

3 Empirical strategy

3.1 Identification strategy

The staggered implementation (over time and across states) of the G8 reform is exploited for identification purposes using a difference-in-differences (DiD) approach. My main DiD model is captured by the equation:

$$zscore_{ist} = \beta_0 + \beta_1 G8_{st} + \alpha X_{ist} + \delta_s + \gamma_t + \varepsilon_{ist}, \tag{1}$$

where $zscore_{ist}$ is the PISA reading, math, or science (standardized) score measured in year t for an academic-track student i in state s. $G8_{st}$ is the G8 reform indicator which equals one if a student observed in year t and in state s belongs to the cohort treated by the G8 reform in that state, and zero otherwise. This is my main variable of interest, as its coefficient β_1 measures the impact of the reform on the treated group after covariates adjustment. X_{ist} is a vector of student and school controls. δ_s and γ_t represent state and cohort fixed effects, respectively. The state (cohort) fixed effects control for unobserved factors that differ across states and not over cohorts (over cohorts and not across states). ε_{ist} is an individual-specific error term.

I also estimate two additional specifications of equation (1), where the G8 dummy is replaced by a variable indicating the duration of treatment, or the average year-week hours of instruction allocated across grades five to nine. The purpose of these additional specifications is to account for cohort-specific treatment intensities.

Several potential threats to internal validity arise when estimating the DiD model just described. The key identifying assumption is, however, that, in the absence of treatment, the difference in outcomes between treatment and comparison groups is constant over time (common trend assumption). Accordingly, a disadvantage of my identification strategy is that any state-specific shock contemporaneous to the G8 reform will bias my estimates. I address this concern in a number of ways in Section 6, where I also run a battery of specification checks with the aim of increasing the confidence in my identification strategy.

3.2 Treatment definitions

Table 1 and Figure 4 define the treatment status of each PISA cohort. Table 1 displays the timing of G8 adoption in each federal state (column 1), the grade(s) initially treated (column 2), and the year of academic-track enrollment (tracking year) for the cohorts attending those grades (column 3). Reported in bold in columns 2 and 3 are those grades and tracking years that, together with the relevant tracking calendars displayed in Figure 4, define the treatment status (T for treatment, C for control) of each PISA cohort, as displayed in columns 4 (2000) to 8 (2012).[14] Finally, treatment status is reported in bold for double cohorts (i.e., the last G9 (**C**) or the first G8 (**T**) cohorts).

Academic-track ninth-graders observed in a treated cohort are assumed to be assigned to G8 since tracking. However, the length of treatment may vary across states and, within a state, across cohorts. For cohorts treated in states where tracking takes place in grade seven (i.e., the PISA 2009 and 2012 cohorts from Berlin and Brandenburg, and the PISA 2012 cohort from Mecklenburg-Vorpommern), or that switched to G8 in grade seven or eight (i.e., the PISA 2006 cohorts from Saxony-Anhalt and Mecklenburg-Vorpommern, respectively), the length of treatment experienced until grade nine is shorter then the modal treatment duration (five years). To capture this heterogeneity in the intensity of treatment, I define a variable indicating the duration of treatment. Moreover, to capture state- and cohort-specific allocation of instruction time, I compute from official historical timetables (KMK, 1997-2014) the average year-week hours of instruction allocated across grades five to nine by state and cohort, and assign this variable to each state-specific cohort in my sample. These variables replace the G8 reform dummy while estimating slightly different specifications of equation (1).

[14] Treatment assignment is somewhat problematic for the PISA 2009 cohort from Hesse. In this state, the G8 reform was introduced for the cohort of 2004 fifth-graders only in 10% of the academic-track high schools. Given the low probability of treatment assignment, I keep Hesse in the sample assuming that its PISA 2009 cohort of ninth-graders – tracked in 2004 – was not affected by the G8 reform. However, I find that my results are not affected by the exclusion of this cohort (or of this state). These results are available upon request.

4 Data

I use data collected in Germany for the first five cycles (2000, 2003, 2006, 2009, and 2012) of the Programme for International Student Assessment (PISA).[15] While the international version of PISA assess 15-year-old students, its German extensions (PISA-E 2000, 2003, and 2006) enlarged the original age-15 PISA samples by collecting additional grade-9 and age-15 samples. In 2009 and 2012, (smaller) grade-9 samples were also collected in addition to the original PISA samples. Because the original age-15 PISA 2009 sample has not been released with a state identifier, I pool grade-9 samples from PISA-E 2000, 2003, and 2006 and from PISA 2009 and 2012.

In each PISA cycle, a range of relevant skills and competencies are assessed in the three domains of reading, mathematics, and science.[16] Each domain is tested using a broad sample of tasks with differing levels of difficulty to represent a coherent and comprehensive indicator of the continuum of students' abilities. Using item response theory, PISA maps performance in each domain on a scale with an international mean of 500 and a standard deviation of 100 test-score points across the OECD countries included in the study. PISA scores are averages of five plausible values, which are drawn from a distribution of values that a student with the given amount of correct answers could achieve as a test score (OECD, 2012).

Depending on the domain tested, my main samples include about 30-34,000 academic-track high school students whose skills were assessed by PISA over the period 2000-2012.[17]

[15] Baumert et al. (2009); Prenzel et al. (2007, 2010); Klieme et al. (2013); Prenzel et al. (2015)

[16] An issue related to the pooled nature of my data regards the comparability of PISA tests across cycles. While reading tests are directly comparable across all cycles, mathematics and science tests underwent major revisions in 2003 and 2006, respectively. However, under the plausible assumption that the degree to which the tests differ is orthogonal to the timing of the introduction of the G8 reform, the DiD estimator employed in this study – which is not a simple before-after estimator, but also takes into account the time trend in the control group – does not require comparability across cycles. In any case, estimating the models on truncated samples – i.e., excluding 2000 and 2000-2003 for math and science, respectively – delivers similar results, available upon request.

[17] More specifically, the domain-specific pooled samples include about 34,000 academic-track students assessed in reading, and about 30,000 students assessed in mathematics (science). This sample size difference is due to the fact that in PISA-E 2000 only about 5/9 of the students assessed in the major domain (reading) were assessed in the other domains through the standard PISA test. Although supplementary mathematics and science national tests were implemented in a second day of testing for all the students assessed in reading, they were based on questions more closely related to German curricula (Stanat et al., 2002). To ensure comparability of test scores across different PISA cycles, I therefore keep in the sample only academic-track students assessed by the standard domain-specific test, and use the domain-specific

Besides student achievement measures, a range of information about the contexts for learning is collected in each PISA cycle by administering background questionnaires to students, parents, teachers, and school principals. Based on questions that are comparable across cycles, two groups of variables are defined at the student and at the school level, and employed as controls in the empirical analysis. Descriptive statistics on these variables are reported in Table 2.

Student controls include a set of demographic and socio-economic characteristics. Among the demographic characteristics, besides a dummy indicating female students and a quadratic age term that controls for potential age/maturation effects, a grade retention dummy is included to control for different schooling experiences.[18] The socio-economic characteristics include an indicator for the number of books at home, a dummy indicating if the student is the only child, two indicators for parents' highest educational level (ISCED), as well as the parents' Highest International Socio-Economic Index (HISEI). There are also variables indicating a student's migration background, namely whether the student was born in a foreign country, whether a foreign language is spoken at home, and whether at least one of the parents was born in a foreign country.

School controls include the total number of enrolled students, the percentage of girls enrolled, the student-teacher ratio, as well as dummy variables indicating urban schools – i.e., schools located in a community of more than 100,000 inhabitants – and privately run schools. Moreover, although PISA does not provide objective measures of the school financial situation, school resources are proxied by the school principals' subjective assessments of whether a lack of instructional material or a lack of computers hindered instruction at their school.

final student weights available for PISA-E 2000 students. In results available upon request, however, I find that my findings are robust to the use of PISA-E 2000 national test scores.

[18]The results reported in Section 7 indicate that the G8 reform did not significantly affect grade retention, hence providing an argument for its inclusion in the main specification. Omitting grade retention delivers, however, similar results, available upon request.

5 Results

5.1 Main Results

The main estimation results are reported in panel A, B, and C of Table 3 for the domains of reading, math, and science, respectively. Each panel row reports coefficients (and standard errors) estimated from separate OLS regressions. The coefficients obtained estimating equation (1) – where the treatment is captured by a G8 reform indicator – are reported in the first row of each panel. The second and third rows of each panel display the coefficients estimated after substituting the G8 reform dummy in equation (1) with alternative measures of treatment intensity: The duration of treatment, and the average year-week hours of instruction allotted across grades five to nine, respectively. Estimation is performed according to the procedure recommended in OECD (2012). For each domain, OLS regressions are run separately on each of the five plausible values,[19] and the results aggregated to obtain the final estimated coefficients and their respective standard errors.[20] Standard errors are clustered on the state level to account for serial error correlation within states over time.[21] In all instances, final sample weights are used to take into account the complex survey nature of PISA data (OECD, 2012).

The results obtained estimating equation (1) in its baseline specification are reported in column (1). Next, to account for compositional changes over time between treatment and control groups, I progressively add the two sets of control variables reported in Table 2: Specification (2) – in column (2) – includes student controls; specification (3) – in column (3) – further adds school controls.[22] Overall, the parameter estimates of the

[19] Plausible values are standardized to have mean zero and variance one in the population of ninth graders from each PISA cycle.

[20] Estimation is performed using the Stata `pv` command.

[21] Although this approach may lead to over-rejection of the null hypotheses when the number of clusters (n) is small (Cameron and Miller, 2015), this does not appear to be an issue in my setting (where $n = 16$ states): The *p-values* obtained from the wild cluster bootstrap procedure (Cameron et al., 2008) provide similar inferential results, available upon request.

[22] As with any survey data set, each PISA sample contains missing values in some background variables (the missing rate is, however, relatively low – generally below five percent – in the pooled sample). This issue is addressed in the empirical analysis by recoding the missing values to zero and including in the estimated models dummy variables indicating the presence of missing values in each of the affected variables when the latter are included in the specification. Similar results, available upon request, were obtained by dropping missing values.

reform effects remain stable across these specifications. This implicitly validates the use of the G8 reform as a quasi-natural experiment, as student and school characteristics that may be correlated with student achievement do not appear to be correlated with the reform, and their omission would not significantly bias its baseline estimated impact. Nonetheless, in order to improve the precision of my estimates, I use specification (3) as the main specification to conduct the remaining empirical analysis.

The coefficients obtained estimating equation (1) in its main specification – first row of column (3) in each panel – indicate that the G8 reform had positive and significant effects on the reading, mathematics, and scientific literacy of academic-track ninth-graders in treated states. In those states, the reform significantly increased PISA standardized scores by a similar order of magnitude: on average, 0.073 standard deviations in reading and mathematics, and 0.087 standard deviations in science. The coefficients estimated on the duration of treatment and on the average year-week hours of instruction allotted across grades five to nine under the main specification – second and third rows of column (3), respectively, in each panel – are in line with the former results. Depending on the domain, each additional year of exposure to treatment led to a significant increase in test scores of 0.013-0.016 standard deviations; or, the modal treatment duration (five years) increased test scores, on average, by 0.065-0.08 standard deviations. Similarly, an additional year-week hour of instruction improved test score by 0.023-0.032 standard deviations. That is, 2.5 additional hours of instruction per week delivered from fifth to ninth grade improved test scores, on average, by 0.058-0.08 standard deviations.

My findings are consistent – although not directly comparable – with those provided by Lavy (2015) and Rivkin and Schiman (2015) using international PISA 2006 and 2009 data, respectively, from a number of OECD countries. These studies exploit within-student variation in subject-specific hours of instruction delivered during the lower secondary school grade attended by 15-year-old students, and find that one additional hour of instruction per week increases test scores, on average, by 0.02-0.06 standard deviations. In contrast, I exploit an increase in average year-week hours of instruction that occurred, for treated cohorts, in the grades attended since academic-track enrollment: An additional

year-week hour of instruction is therefore equivalent to one additional hour of instruction per week delivered during each of the first five high school grades. Compared to the magnitude of this variation, the economic magnitude of the reform effects appears to be small. This is confirmed also by another of the benchmarks suggested by Hill et al. (2008) to examine and interpret effect size measures in education research. In particular, I compare the size of my estimated effects to the typical gains in domain-specific PISA scores from an additional lower secondary school year in Germany. To this end, I pool PISA-E 2000 and 2003 age-15 samples, keeping in the sample students attending eight, ninth, and tenth grade, i.e., the lower secondary school grades attended by the vast majority of 15-year-olds assessed by PISA. For each domain, the annual growth in achievement from a year of schooling is computed as the difference of mean scores in adjacent grades, i.e., grade 8-9 and grade 9-10, respectively – and then converted to a standardized effect size, by dividing it by the pooled student-level standard deviation for the two adjacent grades (Hill et al., 2008). The resulting estimates are reported in Table A.2. Depending on the domain, the typical annual gains from a school year range from 0.47 to 0.56 standard deviations (grade 9-10), or from 0.81 to 0.88 standard deviations (grade 8-9). In contrast, my estimates of the G8 reform effect indicate that an increase of instruction time (across grades five to nine) of about 40 percent of a school year increases average test scores by about 0.073-0.087 standard deviations, depending on the domain. For example, the estimated reform effect for reading (0.073) is only about 8 percent of the typical annual gain in reading scores estimated for grade 8-9 (0.80) and about 15 percent of the gain estimated for grade 9-10 (0.50). Similar patterns are found for mathematics and for science. Thus, my estimated reform effects seem relatively small compared to the typical school year gains obtained in the absence of the reform, although it must be noted that these are gains from a "year of life", capturing also any learning and maturation occurring outside school (Hill et al., 2008).

5.2 Heterogeneity

The estimates reported in Table 3 show the average effects of the reform for the overall population of academic-track ninth-graders, indicating that treated students tend to score significantly better in reading, math, or science tests. However, students' characteristics – such as gender, parental education and migration background, and ability – may affect their capacity to deal with the intensified curriculum introduced by the G8 reform.

To shed further light on the effects of the reform, I estimate additional specifications that explore possible heterogeneous policy effects by adding to the main specification an interaction term between each of the categories considered (gender, parental education and migration background, and grade retention) and the G8 dummy. The coefficients estimated on the reform dummy and on its interaction with the category considered are reported in columns (1) to (4) of Table 4.

The first distinction I consider is gender. As a consequence of behavioral and developmental diversity, boys and girls of the same age may have responded differently to the increased learning intensity introduced by the G8 reform. In particular, girls may have developed a wider set of non-cognitive skills – i.e., attitudes, behaviors, and strategies such as motivation, perseverance, and self-control – that might allow them a better adaptation to the new learning environment (Spinath et al., 2014). The results – reported in column (1) – partially confirm this hypothesis, suggesting that the reform effect in the reading domain is entirely driven by girls. Given that girls in my sample outperform boys in reading even before the introduction of the reform, this finding is consistent with the hypothesis that the effects of a more intensive instruction are heterogeneous based on initial skill differences, with students equipped with higher existing skills benefiting from higher returns (Cuhna and Heckman, 2007). Girls also tend to benefit significantly more than boys in science. In contrast, I do not find evidence of heterogeneous reform effects by gender in math, where boys tend to outperform girls (Fryer and Levitt, 2010).

Further distinctions are by parental education and migration background. The performance of students with less educated parents, or with migrant parents, might have been negatively affected by the reform, possibly because of a lack of parental support in

dealing with the increased pace of learning. However, it may also be the case that those same students benefited from longer school days and/or from increased support from their peer groups. The results – reported in columns (2) and (3) – provide little evidence that the reform significantly enlarged inequality arising from socio-economic background. In contrast, I do find evidence that inequality arising from parental migration background is significantly enlarged by the reform. In particular, compared to students with no parental migration background, students with parental migration background suffer a loss in test scores in all domains, although the loss is not statistically significant at standard levels in reading.

Finally, in column (4) I consider heterogeneous reform effects by grade retention. The latter can be viewed as a low achievement (low ability) proxy. Low achievers are particularly vulnerable to the reform as they are most at risk of experiencing difficulties in adjusting to the new learning environment. It is reasonable to expect the effects of a more intensive instruction to be heterogeneous based on initial skill differences, with the most harmful – or less beneficial – effects on the students with lower existing skills, i.e., those that benefit from lower returns on the existing skills (Cuhna and Heckman, 2007). The evidence is consistent with this hypothesis: The estimated differential reform effect for low achievers is negative and significant in all domains, suggesting that the average reform effects are essentially driven by high achievers. Compared to high achievers, low achievers experience a significant loss in test scores ranging from 0.12 standard deviations (mathematics) to 0.166 standard deviations (science). This finding points to important heterogeneous reform effects, as low achieving students appear to be less capable of coping with the higher per-grade curriculum requirements introduced by the G8 reform. As previously discussed, I investigate this issue in Andrietti and Su (2016), finding evidence consistent with this hypothesis.

6 Robustness

In this section, I address the main concerns that might threaten my identification strategy. First, I present – in Figure 5 – graphical evidence in support of the common trend assumption. Then, I provide further – regression based – evidence supporting the common trend assumption, and assess the sensitivity of my results to multiple robustness checks, demonstrating that the G8 reform effects are very similar across different specifications. The results of the robustness analysis are reported in columns (2) to (11) of Table 5, where, for the sake of comparison, column (1) reports the results obtained estimating equation (1) in its main specification, i.e., including student and school controls.

6.1 Common trends

The key identification assumption behind the DiD approach is that treatment and comparison groups follow a common trend in the absence of the reform, i.e., there are no unobserved variables that change over time resulting in differential effects on test scores of students that were treated by the G8 reform and students that were not. Equivalently, the treatment must be the only reason why treatment and control group trends deviate in the post-reform period. The main concern is therefore that the reform effects reflect differential time trends in the outcomes of interest between treatment and comparison states, rather than a true policy impact. Below, I address this concern in a number of ways.

6.1.1 Inter-temporal reform effects

While a direct test of the common trend assumption is not possible, given the unobservability of the treatment counterfactual, graphical and regression based evidence might be used to corroborate its validity. In Figure 5, I present point estimates (with 95 percent confidence intervals) from baseline regressions designed to capture inter-temporal reform effects.[23] The baseline specification includes – besides state and time fixed effects – an

[23] Anderson and Walker (2015) use this approach to analyze the effect of shortening the school week on student performance.

indicator of the first state-specific G8 cohorts observed during my sample period, as well as three lead indicators and two lag indicators. The lead dummies take on a value of one for the cohorts assessed three, six, or nine years prior to the first G8 cohorts observed in my sample, respectively. The lag dummies takes on a value of one for cohorts assessed three (six) years after the first G8 cohorts. The omitted category is represented by the cohorts assessed twelve years prior to the first G8 cohorts observed in my sample. The pattern of inter-temporal reform effects is consistent with the common trend assumption. The coefficients estimated for the lead dummies are both economically and statistically insignificant in all the domains, indicating that students in states that switched to the G8 regime share similar pre-treatment trends in test scores with students in states that remained in the G9 regime. In contrast, the first G8 cohorts observed in my sample experience a sharp increase in standardized test scores in all domains. Importantly, this improvement persists over time in all the domains tested, and becomes statistically significant for science.

6.1.2 Placebo treatments

A simple way to enhance the graphical evidence displayed in Figure 5 is a placebo treatment test in the years preceding the actual treatment that can show deviations from the common trend in pre-treatment years. I run this test by including in equation (1) a placebo reform dummy indicating the state-specific cohorts that immediately precede the first G8 cohorts observed in my sample. A significant estimated coefficient on this dummy would indicate different trends in outcomes for treatment and control groups before the G8 reform actually kicked in. However, consistent with the graphical evidence presented in Figure 5, this coefficient – in column (2) – turns out to be close to zero and insignificant in all the domains. Furthermore, I provide an additional placebo test, based on the idea that the achievement of basic- and middle-track students in treated states should not be significantly affected by the G8 reform, as they were not directly exposed to it. The insignificance of the G8 dummy estimated coefficient – in column (3) – also confirms this expectation. Taken together, the evidence proceeding from these falsification tests and

from Figure 5 corroborate the validity of the common trend assumption.

6.1.3 State-specific shocks

A drawback of the DiD approach is that it does not control for state-specific shocks, which might similarly affect all students in a state, for example, due to changes in primary school. One way to address this concern is by allowing for state-specific linear time trends. The idea is to use the pre-reform data to extrapolate the time trend of each state into the post-reform periods. This allows treatment and comparison states to follow different secular trends in a limited but potentially revealing way (Angrist and Pischke, 2009). Besley and Burgess (2004) show that allowing for differential time trends in a DiD regression may destroy otherwise large and statistically significant treatment effects. It is therefore reassuring that my main results are robust to the inclusion – in column (4)– of state-specific linear trends.

6.1.4 Difference-in-difference-in-differences

As an alternative, and more flexible, way to control for both state-specific trends and regional shocks potentially correlated with the G8 policy, I exploit the fact that the latter was implemented at different points in time across different states and affected academic-track students but not middle-track students. Adding middle-track students as an additional control group leads to a difference-in-difference-in-differences (DDD) model that makes use of the outcome change of middle-track students to control for state-specific shocks potentially correlated with the policy.[24]

The model is captured by the following baseline equation:

$$zscore_{iast} = \beta_0 + \beta_1 G8_{st} + \beta_2 Atrack_{ist} + \beta_3 G8_{st} \times Atrack_{ist} + \delta_{sa} + \gamma_{ta} + \lambda_{st} + \varepsilon_{iast}, \quad (2)$$

[24] Besley and Case (2000) discuss the conditions under which DiD and DDD estimators deliver unbiased estimates, emphasizing that the latter are crucially dependent on the quality of the control group chosen. In the German three-track educational system, middle-track students represent, among the students that were not affected by the G8 reform, the group that is most closely comparable to academic-track students. It seems therefore plausible to assume that academic- and middle-track students are comparable, i.e., respond similarly to state-specific shocks.

where s indexes state, t indexes time (cohort), and a indexes track. $Atrack_{ist}$ is a dummy taking the value 1 for academic-track students in state s and time t, and 0 for middle-track students. The parameters δ_{sa}, γ_{ta}, and λ_{st} are, respectively, state-by-track, time-by-track, and state-by-time fixed effects.[25] The state-by-track effects account for state-specific factors that vary across tracks but are fixed over time. These include, for example, fixed-differences across states in terms of educational policies and local labor market opportunities. The time-by-track effects account for time varying and track-specific factors that are common across states. The state-by-time effects account for time-varying state-specific factors that have a common effect across tracks. In its main specification, the model also includes a vector of student and school controls, as well as its interaction with the academic-track dummy. The coefficient β_3 represents the impact of the G8 reform on the achievement of academic-track students versus middle-track students in treated states relative to control states. The results obtained from equation (2) – in column (5) –, although estimated less precisely, confirm my main finding. The similarity of the DDD results – in terms of economic magnitude – to the main DiD findings lends further credibility to a causal interpretation of the latter. Moreover, the interpretation of the DDD coefficient of interest as a causal effect relies on a weaker assumption: in the absence of the reform, the difference in outcomes between academic- and middle-track students would have developed similarly in treated and control states. Nonetheless, as this assumption is not testable, I carry out a placebo test similar in spirit to the one carried out in the DD setting, by adding to equation (2) a G8 dummy lead indicator and its interaction with the academic-track dummy dummy. The coefficient estimated on the latter term – in column (6) – is insignificant in all the domains, suggesting that the reform effects estimated with DDD are not confounded by systematic differences in trends between treatment and comparison groups.

6.1.5 Compositional changes

More generally, my identification strategy would be threatened if compositional changes over time were induced by the reform. For example, the distribution of students across school within a German state might have changed in response to the introduction of the reform. Also, the reform could have affected high school grade retention; or, more teachers could have been hired to compensate for the instruction time and curricular compression. One way to address this

[25] Note that state and time fixed effects included in the DDD model are now absorbed by the vector of state-specific time effects, λ_{st}.

concern is to check if student and school controls included in my main specification were affected by the treatment. If this were the case, they may capture part of or bias the treatment effect. The results of estimating equation (1) in its baseline version with student and school controls as dependent variables – reported in Tables A.3 and A.4 – confirm that the reform did not induce compositional changes either at the student- or at the school-level. Furthermore, in Section 7 I show that the reform did not affect grade retention in high school.

Compositional changes might arise from self-selection and/or from non-compliance issues. First, the distribution of students across school types within a German state might have changed in response to the G8 reform. Weaker students that would have enrolled in the academic track offered by comprehensive schools could easily avoid the reform by switching to a lower track within the same school. Or, weaker students that would have enrolled in the academic track in the pre-reform period might rather prefer to enroll in other secondary schools (either lower tracks or comprehensive schools) after the reform. In both cases, I might find a positive reform effect even if the reform had no direct effect on student achievement. While the former case was addressed at the sample selection stage, excluding from the sample academic-track students enrolled in comprehensive schools, I address the latter possibility estimating equation (1) with academic-track attendance (vs. attendance of other types of secondary schools) as dependent variable. The statistical and economic insignificance of the G8 reform coefficient – in column (7) – suggests that the reform effects do not proceed from a change in the distribution of students across school types, and is consistent with the earlier findings of absence of compositional changes induced by the reform. As a consequence, selection out of the sample – i.e., into other secondary school tracks – should not be a major concern here.

Moreover, since the reform was introduced in an entire state at one time, avoiding the reform while staying in the academic track – i. e., self-selecting into the control group – would require moving to a different state, an unlikely possibility considering the high costs associated with residential mobility. A more plausible scenario is that students from treated states living at the border of control states would avoid the reform by attending high school in the control states. While PISA data do not offer information on student residence, it is likely that the number of cross-border commuters is very small.

Further concerns arise from non-compliance to the treatment that might have affected those states where the G8 reform was announced – and therefore anticipated – before its implemen-

tation. Although in principle students in the first G8 cohorts or in the last G9 cohort might have tried to switch to G9 – by skipping a grade – or to G8 – by voluntary repeating a grade, respectively, it is very unlikely that they actually did so, as in either case they would end up graduating in their original cohort. Moreover, these concerns only apply to students belonging to the double cohort. Later, I will assess whether my results are driven by the peculiarity of this cohort.

6.1.6 Contemporaneous policy changes

My DD results could also be biased by contemporaneous policy changes. A reform of the German high school system that directly affected the academic-track (as well as the lower-tracks) is the introduction of Centralized Exit Examinations (CEEs). While CEEs were introduced long before the start of my observation period in some federal states, most of the remaining states introduced CEE between 2005 and 2008. Jürges et al. (2012) provide evidence that CEEs do not matter significantly either for students in academic-track or for literacy skills tests like the ones analyzed in this study. However, to allow for the possibility that the introduction of CEEs affected students exposed and not exposed to the G8 reform in different ways, I add to the main specification a dummy capturing the switch to CEEs of some states (i.e., Berlin, Brandenburg, Bremen, Hamburg, Hesse, Lower Saxony, North Rhine-Westfalia, and Schleswig Holstein) during my observation period. The results of estimating this model – displayed in column (8) – show that the G8 effects, despite being estimated somewhat less precisely, preserve their economic magnitude for all domains and their statistical significance for reading and science.

6.2 Further sensitivity analysis

6.2.1 Double cohorts

The first cohort that experienced the G8 regime in each state was considered part of a *double graduating cohort* because it was expected to graduate at the same time as the last cohort graduating under the G9 regime. Each double graduating cohort was approximately twice as large as earlier or later cohorts and was therefore domain to much stronger competition for post-graduation resources (jobs, admission to university degree programs, etc.).[26] Anecdotal

[26]See Morin (2015a) and Morin (2015b) for an analysis of the effects of the increased competition arising from Ontario's double cohort on the earnings of high school graduates, and on university grades,

evidence says that parents were worried about the consequences on the future academic and labor market outcomes of their children possibly deriving from the increased competition. At the same time, it might be that students experienced this increasing pressure as an incentive to work harder. The increased competition/pressure should not be a major cause of concern in my setting, given that the first treated cohorts assessed by PISA at the end of their ninth grade are still three years apart from graduation. However, it might also be the case that teachers or schools reacted to the reform by reallocating efforts toward treated cohorts and away from non-treated cohorts, when both treated and non-treated cohorts attended the same schools at the same time (i.e., for double cohorts). It is therefore interesting to check whether my results are driven by these double cohorts. To this end, I add to my main specification a dummy variable which equals one for the double cohorts (i. e., either the last G9 cohort or the first G8 cohort, as indicated in Table 1) observed in my sample. Estimating the model under this specification – in column (9) – confirms that my main results are not driven by peculiarities pertaining to the double cohorts.

6.2.2 Alternative samples

Finally, I assess the robustness of my results to the use of two alternative samples. First, I exclude from the main samples those states whose cohorts were either tracked in grade 7 at some point and/or were partially treated. The results – reported in column (10) – are qualitatively the same, despite being estimated with less precision. Second, given that the source of my identification are states whose observed cohorts switched at some point to G8, my results should not be affected by the exclusion of states whose cohorts were always treated during my observation period (i.e., Saxony and Thuringia). This expectation is confirmed by the results reported in column (11).

7 High school grade retention and remedial education

Grade retention and remedial education represent important costs to the educational system. They may also serve as indicators of the student ability to deal with the increased learning intensity introduced by the G8 reform. It is therefore important to provide additional pieces

respectively.

of evidence on possibly unintended effects of the reform, by documenting its effects on the probability of repeating a high school grade or of participating in remedial education. To this end, I estimate linear probability DiD models.

The grade retention model is represented by the following equation:

$$Repeat\_high_{ist} = \beta_0 + \beta_1 G8r_{st} + \alpha X_{ist} + \delta_s + \gamma_t + \varepsilon_{ist}, \tag{3}$$

where $Repeat\_high_{ist}$ equals one if a student experienced grade retention during high school, and zero otherwise, and $G8r_{st}$ equals one if a student entered high school *after the first treated cohort*, and zero otherwise. In states where multiple grades switched to G8 at the same time, i.e., Bavaria, Mecklenburg-Vorpommern, Niedersachsen, and Saxony-Anhalt, the first treated cohort corresponds to the highest grade initially treated. In contrast, the first treated PISA cohorts that I observe in these states – PISA 2006 in Saxony-Anhalt and Mecklenburg-Vorpommern; PISA 2009 in Bavaria and Lower Saxony – do not correspond to the highest grade initially treated. In Saxony-Anhalt and Mecklenburg-Vorpommern, the G8 reform – introduced in 2003 and 2004, respectively – affected grades five to nine. This means that PISA 2006 academic-track ninth-graders switched to G8 when they were in grade seven (Saxony-Anhalt) or eight (Mecklenburg-Vorpommern), and that earlier cohorts were treated since their eight or ninth grade. Similarly, in Bavaria and Lower-Saxony, the reform – introduced in 2004 – affected contemporaneously grades five and six. Thus, although the PISA 2009 cohorts in Bavaria and Lower Saxony were treated since grade five, an earlier cohort was treated since grade six. Assigning these cohorts to treatment might therefore be problematic because high school grade retention found in these cohorts could have happened in grades that were not yet exposed to treatment (i.e., grades five to seven in Mecklenburg-Vorpommern, grades five and six in Saxony-Anhalt, and grade five in Bavaria and Lower-Saxony). To avoid this contamination issue, I drop these cohorts from the sample.

I first estimate equation (3) in its main specification. Then, I estimate additional specifications that explore possible heterogeneous policy effects by gender, parental education, and parental migration background. Table 6 reports the estimated coefficients on the reform dummy and on its interaction with the category considered. The first finding – in column (1) – is that the G8 reform has no effect on the probability of repeating a grade in high school. This is consistent

with the evidence provided by Huebener and Marcus (2015), based on administrative data, that the reform did not affect repetition rates in grades seven to nine. However, the heterogeneity analysis – in columns (2) to (4) – reveals that the probability of high school grade retention is significantly higher after the reform for students with parental migration background. This result is consistent with the heterogeneous reform effects by parental migration background found for cognitive skills, providing further evidence that the reform enlarged the migration background gap in student achievement.

The remedial education model is represented by the following equation:

$$Remedial_{ist} = \beta_0 + \beta_1 G8_{st} + \alpha X_{ist} + \delta_s + \gamma_t + \varepsilon_{ist}, \tag{4}$$

where $Remedial\_ist$ equals one if a student i in state s is taking remedial education courses at time t, and zero otherwise. Questions on remedial education were asked in PISA 2000, 2003, 2009, and 2012. However, the domain focus changed over time and/or across margins (extensive vs. intensive). For example, while PISA 2000, 2009, and 2012 only asked about participation in remedial courses (in German and overall, in each domain, and in German only, respectively), PISA 2003 also asked about the time spent in remedial education (intensive and, implicitly, extensive margin) in math and overall. I therefore focus on the extensive margin observed at least in a pre- and a post-reform period. The probability of participating in remedial education in German, math, or in any domain is estimated on pooled PISA samples from 2000, 2009, and 2012, from 2003 and 2009, and from 2000, 2003, and 2009, respectively. The results obtained estimating equation (4) in its main specification – in columns (5) - (7) of Table 6 – indicate that the reform did not affect the probability of remedial education in German. However, treated students are significantly less likely to participate in remedial education courses in math or in any domain. The latter finding should be, however, interpreted with caution. While at first blush it may suggest that the reform reduced the need of remedial work, it may also be the case that the reform simply crowded out any remedial work, which may be still much needed, due to the time commitment of additional instruction hours coring new learning content.

8 Conclusion

Time of instruction is an intuitive, and yet understudied, input in educational production. Most well identified studies exploit short-lived and sometimes targeted or bundled policies that alter instruction time while keeping unaltered the intensity of learning. In contrast, I exploit a universal educational policy (the G8 reform in Germany) that compressed academic-track high school curriculum into a (one-year) shorter time span, and induced a large and lasting increase in instruction time and in the intensity of learning, without directly affecting other school inputs.

Using 2000-2012 PISA data and a quasi-experimental approach, I estimate the impacts of the intensified curriculum introduced by the G8 reform on cognitive skills. I find robust evidence that the reform improved, on average, the reading, mathematical, and scientific literacy skills acquired by academic-track high school students upon treatment. However, the effects appears to be small and heterogeneous across gender, parental migration background, and student ability. In particular, inequalities arising from parental migration background and from student ability appear to be significantly enlarged by the reform. Moreover, although I find little evidence of possible unintended reform effects on other relevant margins (high school grade retention, remedial education), I do find evidence that high school grade retention increased for students with parental migration background. Taken together, these results suggest that the effects of a more intensive instruction might be heterogeneous based on initial skill differences, with the most harmful – or less beneficial – effects on the students with lower existing skills, i.e., those that benefit from lower returns on the existing skills. In Andrietti and Su (2016) I find evidence consistent with this hypothesis: While the G8 reform improves student test scores on average, such a benefit is much more pronounced for well-prepared students; in contrast, less prepared students do not benefit from the reform.

Given the importance of cognitive skills – as standardized measures of short-term student achievement – for longer-term economic outcomes (Hanushek and Wössmann, 2008, 2011, 2012, 2015), my findings offer important policy insights. A major issue of the public debate over the G8 reform in Germany concerns the question of whether it is possible to improve educational performance by increasing the learning intensity in high school. Based on fears that the intensified curriculum introduced by the G8 reform will overburden students, thereby negatively affecting their educational achievement, some states are considering, or have already implemented, a (partial) switch back to the old regime. My findings suggest that moving to a compressed

high-school curriculum did not compromise and benefited, on average, students' cognitive skills. However, the enlarged gaps in student achievement discussed above also raise equity concerns that policy-makers should be aware of. These concerns may be particularly important for Germany, given the changes in demographics (e.g., a rapid increase in the share of the immigrant population) that the country is currently experiencing.

Compared to earlier G8 studies, the internal validity of my findings is improved by assessing their robustness to the use of a stronger identification strategy, i.e., a triple-difference approach that confirms the double-difference results. The external validity is also improved, at least within Germany, by the use of a more representative dataset, and the focus on a broader set of outcomes for students still in school. However, generalizing my findings to other educational systems requires caution, given the specificities of the German educational context, where students are tracked into differing-ability schools as early as at age 10. While my results may be relevant to countries that similarly track students across schools (e.g., Austria and, to a lesser extent, Hungary and Slovakia), differences in timing and type of tracking could limit a broader generalizability of my results. On the one hand, the evidence on increase inequality in achievement caused by early tracking (Hanushek and Wössmann, 2006) suggests that the equity concerns raised by my results may be less relevant to countries where students are tracked later in their schooling career, either across schools or within schools, or to countries that do not use ability-tracking. On the other hand, the benefits of additional instruction time might be lower and/or more heterogeneous in systems without tracking, because of more heterogeneous peer environments.

The educational policy change introduced by the G8 reform is appealing because overall class hours are not increased, so no new resources are required. However, to the best of my knowledge, universal policy reforms that would similarly compress learning time or influence learning intensity have not been implemented (or considered) yet in other countries.[27] Perhaps as a consequence, this is the first study to look at the effects of this type of reform on short-term standardized measures of student achievement.

Beyond its policy relevance, my study contributes to the literature on the role of instruction time as a school input by showing that students might benefit from increased time of instruction time despite the increased burden of a higher intensity of learning. Although the benefits appear

[27] Although in the Canadian province of Ontario high school duration was recently shortened by one year, the reduction was compensated by a corresponding cut in the overall curriculum. Therefore, the share of the overall curriculum covered per grade was not affected, and the intensity of learning remained unaltered.

to be small – perhaps just because of the increased intensity of learning – they are consistent with recent studies that exploit similar sources of exogenous variation (Lavy, 2012; Fryer, 2014), or particular data features (Lavy, 2015; Rivkin and Schiman, 2015), to overcome biases introduced by the non-random allocation of instructional time. These various estimates suggest that some consensus is being reached over the nature of the causal relationship between instructional time and student achievement. In particular, my results suggest that this relationship holds even in settings where the intensity of learning is increased.

References

Agüero, Jorge M., and Teresa Beleche (2013) 'Test-mex: Estimating the effects of school year length on student performance in mexico.' *Journal of Development Economics* 103, 353–316

Allensworth, Elaine, Takako Nomi, Nicholas Montgomery, and Valerie E. Lee (2009) 'College preparatory curriculum for all: Academic consequences of requiring algebra and english i for ninth graders in chicago.' *Educational Evaluation and Policy Analysis* 31(4), 367–391

Anderson, D. Mark, and Mary Beth Walker (2015) 'Does shortening the school week impact student performance? evidence from the four-day school week.' *Education Finance and Policy* 10(3), 314–349

Andrietti, Vincenzo (2015) 'The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment.' UC3M WP Economic Series 15-06, Universidad Carlos III de Madrid, June

Andrietti, Vincenzo, and Xuejuan Su (2016) 'Education curriculum and student achievement: Theory and evidence.' UC3M WP Economic Series 16-07, Universidad Carlos III de Madrid, April

Angrist, Joshua D., and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press)

Aucejo, Esteban M., and Teresa Foy Romano (2014) 'Assessing the effect of school days and absences on test score performance.' CEP Discussion paper 1302, LSE Center for Economic Performance, September

Baumert, J., C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, and M. Weiss (2009) *Programme for International Student Assessment 2000 (PISA 2000). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2000_v1)

Bellei, Cristian (2009) 'Does lengthening the school day increase students' academic achievement? results from a natural experiment in chile.' *Economics of Education Review* 28, 629–640

Besley, Timothy, and Anne Case (2000) 'Unnatural experiments? estimating the incidence of endogenous policies.' *Economic Journal* 110(november), F672–F694

Besley, Timothy, and Robin Burgess (2004) 'Can labor regulation hinder economic performance? evidence from india.' *Quarterly Journal of Economics* 119(1), 91–134

Büttner, Bettina, and Stephan L. Thomsen (2015) 'Are we spending too many years in school? causal evidence of the impact of shortening secondary school duration.' *German Economic Review* 16(1), 65–86

Cameron, Colin A., and Douglas L. Miller (2015) 'A practitioner's guide to cluster-robust inference.' *Journal of Human Resources* 50(2), 317–372

Cameron, Colin A., Jonah G. Gelbach, and Douglas L. Miller (2008) 'Bootstrap-based improvements for inference with clustered errors.' *Review of Economics and Statistics* 90(3), 414–427

Carlsson, Magnus, Gordon B. Dahl, Björn Öckert, and Dan-Olof Rooth (2015) 'The effect of schooling on cognitive skills.' *Review of Economics and Statistics* 97(3), 533–547

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2015) 'The aftermath of accelerating algebra: evidence from district policy initiatives.' *Journal of Human Resources* 50(1), 159–188

Cortes, Kalena E., and Joshua S. Goodman (2014) 'Ability-tracking, instructional time, and better pedagogy: the effect of double-dose algebra on student achievement.' *American Economic Review: Papers & Proceedings* 104(5), 400–405

Cortes, Kalena E., Joshua S. Goodman, and Takako Nomi (2015) 'Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra.' *Journal of Human Resources* 50(1), 108–158

Cuhna, Flavio, and James J. Heckman (2007) 'The technology of skill formation.' *the American Economic Review* 97(2), 31–47

Dahmann, Sarah (2015) 'How does education improve cognitive skills? instructional time versus timing of instruction.' SOEP papers on Multidisciplinary Panel Data Research 769, DIW Berlin, July

Dahmann, Sarah, and Silke Anger (2014) 'The impact of education on personality. evidence from a german high school reform.' SOEP papers on Multidisciplinary Panel Data Research 658, DIW Berlin, May

Dobbie, Will, and Roland Fryer (2013) 'Getting beneath the veil of effective schools: Evidence from new york city.' *American Economic Journal: Applied Economics* 5(4), 28–60

Dörsam, Michael, and Verena Lauber (2015) 'The effect of a compressed high school curriculum on university performance.' Technical Report, University of Kostanz

Dougherty, Shaun, Joshua Goodman, Darryl Hill, Erica Litke, and Lindsay C. Page (2015) 'Early math coursework and college readiness: Evidence from targeted middle school math acceleration.' NBER Working Paper 21395, National Bureau of Economic Research, July

Dougherty, Shaun M. (2015) 'Bridging the discontinuity in adolescent literacy? mixed evidence from a middle grades intervention.' *Education Finance and Policy* 10(2), 157–192

Fitzpatrick, Maria D., David Grissmer, and Sarah Hastedt (2011) 'What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment.' *Economics of Education Review* 30, 267–279

Fryer, Roland G. (2014) 'Injecting charter school best practices into traditional public schools: Evidence from field experiments.' *The Quarterly Journal of Economics* 129(3), 1355–1407

Fryer, Ronald G., and S. D. Levitt (2010) 'An empirical analysis of the gender gap in mathematics.' *American Economic Journal: Economic Policy* 2(2), 210–240

Goodman, Joshua (2014) 'Flaking out: Student absences and snow days as disruptions of instructional time.' NBER Working Paper 20221, National Bureau of Economic Research, June

Hansen, Benjamin (2011) 'School year length and student performance: Quasi-experimental evidence.' Working paper, SSRN http://dx.doi.org/10.2139/ssrn.2269846

Hanushek, Eric A. (2015) 'Time in education: Introduction.' *The Economic Journal* 125(588), F394–F396

Hanushek, Eric A., and Ludger Wössmann (2006) 'Does educational tracking affect performance and inequality? differences-in-differences evidence across countries.' *Economic Journal* 116(510), C63–C76

_ (2008) 'The role of cognitive skills in economic development.' *Journal of Economic Literature* 46(3), 607–668

_ (2011) 'The economics of international differences in educational achievement.' In *Handbook of the Economics of Education,* ed. Stephen Machin Eric A. Hanushek and Ludger Wössmann, vol. 3 (Elsevier) pp. 89–200

_ (2012) 'Do better schools lead to more growth? cognitive skills, economic outcomes, and causation.' *Journal of Economic Growth* 17(4), 267–321

_ (2015) *The knowledge capital of nations: Education and the economics of growth* (MIT Press)

Heckman, James J., Jora Stixrud, and Sergio Urzua (2006) 'The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior.' *Journal of Labor Economics* 24(3), 411–482

Hill, Carolyn J., Howard S. Bloom, Alison R. Black, and Mark W. Lipsey (2008) 'Empirical benchmarks for interpreting effect sizes in research.' *Child development perspectives* 2(3), 172–177

Huebener, Mathias, and Jan Marcus (2015) 'Moving up a gear: the impact of compressing instructional time into fewer years of schooling.' Discussion paper 1450, DIW Berlin, February

Huebener, Mathias, Susanne Kuger, and Jan Marcus (2016) 'Increasing instruction hours and the widening gap in student performance.' Discussion paper 1561, DIW Berlin, March

Jacob, Brian A., Lars Lefgren, and David P. Sims (2010) 'The persistence of teacher-induced learning.' *Journal of Human Resources* 45(4), 915–943

Jürges, Hendrik, Kerstin Schneider, Martin Senkbeil, and Klaus Carstensen (2012) 'Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy.' *Economics of Education Review* 31, 56–65

Klieme, E., C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, and P. Stanat (2013) *Programme for International Student Assessment 2009 (PISA 2009). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2009_v1)

KMK (1997-2014) *Wochenpflichtstunden der Schülerinnen und Schüler - Statistiks 1997 bis 2014*

Kraft, Matthew A. (2015) 'How to make additional time matter: Integrating individualized tutorials into an extended day.' *Education Finance and Policy* 10(1), 81–116

Krashinsky, Harry (2014) 'How would one extra year of high school affect academic performance in university? evidence from an educational policy change.' *Canadian Journal of Economics* 47(1), 70–97

Kühn, Svenja M., Isabell van Ackeren, Gabriele Bellenberg, Christian Reintjes, and Grit im Brahm (2013) 'Wie viele schuljahre bis zum abitur? eine multiperspektivische standortbestimmung im kontext der aktuellen schulzeitdebatte.' *Zeitschrift für Erziehungswissenschaft* 16, 115–136

Lavy, Victor (2012) 'Expanding school resources and increasing time on task: Effects of a policy experiment in israel on student academic achievement and behavior.' NBER Working Paper 18369, National Bureau of Economic Research, September

_ (2015) 'Do differences in school's instruction time explain international achievement gaps in maths, science and language? evidence from developed and developing countries.' *Economic Journal* 125(588), F397–F424

Lee, Jong-Wha, and Robert Barro (2001) 'Schooling quality in a cross-section of countries.' *Economica* 68(272), 465–488

Lohmar, Brigitte, and Thomas Eckhardt (2010) *The education system in the Federal Republic of Germany 2008: A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe* (Bonn: Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs)

Marcotte, Dave E. (2007) 'Schooling and test scores: A mother natural experiment.' *Economics of Education Review* 26, 629–640

Marcotte, Dave E., and Steven W. Hemelt (2008) 'Unscheduled school closing and student performance.' *Education Finance and Policy* 3(3), 316–338

Meyer, Tobias, and Stephan L. Thomsen (2016) 'How important is secondary school duration for post-school education decisions? evidence from a natural experiment.' *Journal of Human Capital* 10(1), 67–108

Meyer, Tobias, Stephan L. Thomsen, and Heidrun Schneider (2015) 'New evidence on the effects of the shortened school duration in the german states: An avaluation of post-secondary education decisions.' IZA DP 9507, Institute for the Study of Labor, November

Morin, Louis-Philippe (2013) 'Estimating the benefit of high school for university-bound students: Evidence of subject-specific human capital accumulation.' *Canadian Journal of Economics* 46(2), 441–468

_ (2015a) 'Cohort size and youth earnings: Evidence from a quasi-experiment.' *Labour Economics* 32, 99–111

_ (2015b) 'Do men and women respond differently to competition? evidence from a major education reform.' *Journal of Labor Economics* 33(2), 443–491

Nomi, Takako, and Elaine Allensworth (2009) "'double-dose" algebra as an alternative strategy to remediation: Effects on students' academic outcomes.' *Journal of Research on Educational Effectiveness* 2(2), 111–148

Nomi, Takako, and Elaine Allensworth (2013) 'Sorting and supporting: why double-dose algebra led to better test scores but more course failures.' *American Educational Research Journal* 50(4), 756–788

OECD (2003) *Literacy Skills for the World of Tomorrow. Further Results from PISA 2000* (OECD Publishing)

_ (2012) *PISA 2009 Technical Report* (OECD Publishing)

Parinduri, Rasyad A. (2014) 'Do children spend too much time in schools? evidence from a longer school year in indonesia.' *Economics of Education Review* 41, 89–104

Pischke, Jörn-Steffen (2007) 'The impact of length of the school year on student performance and earnings: Evidence from the german short school years.' *Economic Journal* 117(523), 1216–1242

Prenzel, M., C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, and R. Pekrun (2010) *Programme for International Student Assessment 2006 (PISA 2006). Version: 1* (IQB - Institut zur Qualitätentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2006_v1)

Prenzel, M., C. Sälzer, E. Klieme, O. Köller, J. Mang, J.-H. Heine, A. Schiepe-Tiska, and K. Müller (2015) *Programme for International Student Assessment 2012 (PISA 2012). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2012_v1)

Prenzel, M., J. Baumert, W. Blum, R. Lehmann, D. Leuner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, and U. Schiefele (2007) *Programme for International Student Assessment 2003 (PISA 2003). Version: 1* (IQB - Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2003_v1)

Rivkin, Steven G., and Jeffrey C. Schiman (2015) 'Instruction time, classroom quality, and academic achievement.' *Economic Journal* 125(588), F425–F448

Sims, David P. (2008) 'Strategic responses to school accountability measures: It's all in the timing.' *Economics of Education Review* 27, 58–68

Spinath, Birgit, Christine Eckert, and Ricarda Steinmayr (2014) 'Gender differences in school success: what are the roles of students' intelligence, personality, and motivation.' *Educational Research* 56(22), 230–243

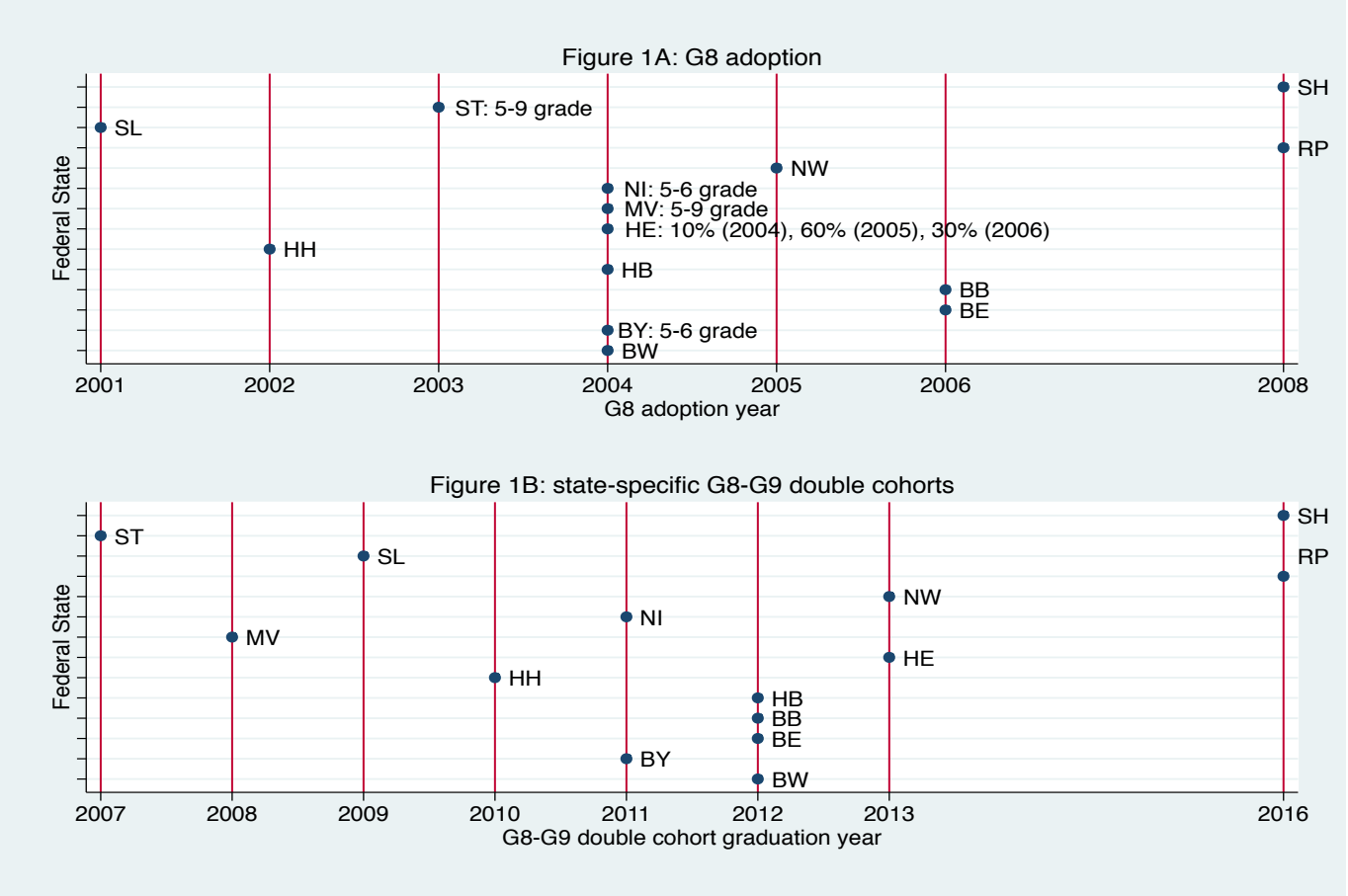
Stanat, P., J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, G. Schümber, K.-J. Tillmann, and D. Weiss (2002) *PISA 2000: Overview of the Study* (Max Planck Institute for Human Development, Berlin)

Taylor, Eric (2014) 'Spending more of the school day in math class: Evidence from a regression discontinuity in middle school.' *Journal of Public Economics* 117, 162–181

Thile, Hendrik, Stephan L. Thomsen, and Bettina Büttner (2014) 'Variation of learning intensity in late adolescence and the effect on personality traits.' *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 177(4), 861–892

Wössman, Ludger (2003) 'Schooling resources, educational institutions and student performance: The international evidence.' *Oxford Bullettin of Economics and Statistics* 65(2), 117–170

Fig. 1. Timing of the G8 reform implementation



Legenda
BW: Baden-Württemberg
BY: Bavaria
BE: Berlin
BB: Brandenburg
HB: Bremen
HH: Hamburg
HE: Hessen
MV: Mecklenburg-Vorpommern
NI: Lower Saxony
NW: North Rehin-Westfalia
RP: Rheinland-Palatinate
SL: Saarland
ST: Saxony-Anhalt
SH: Schleswig-Holstein

Fig. 2. Map of the G8 reform implementation timing

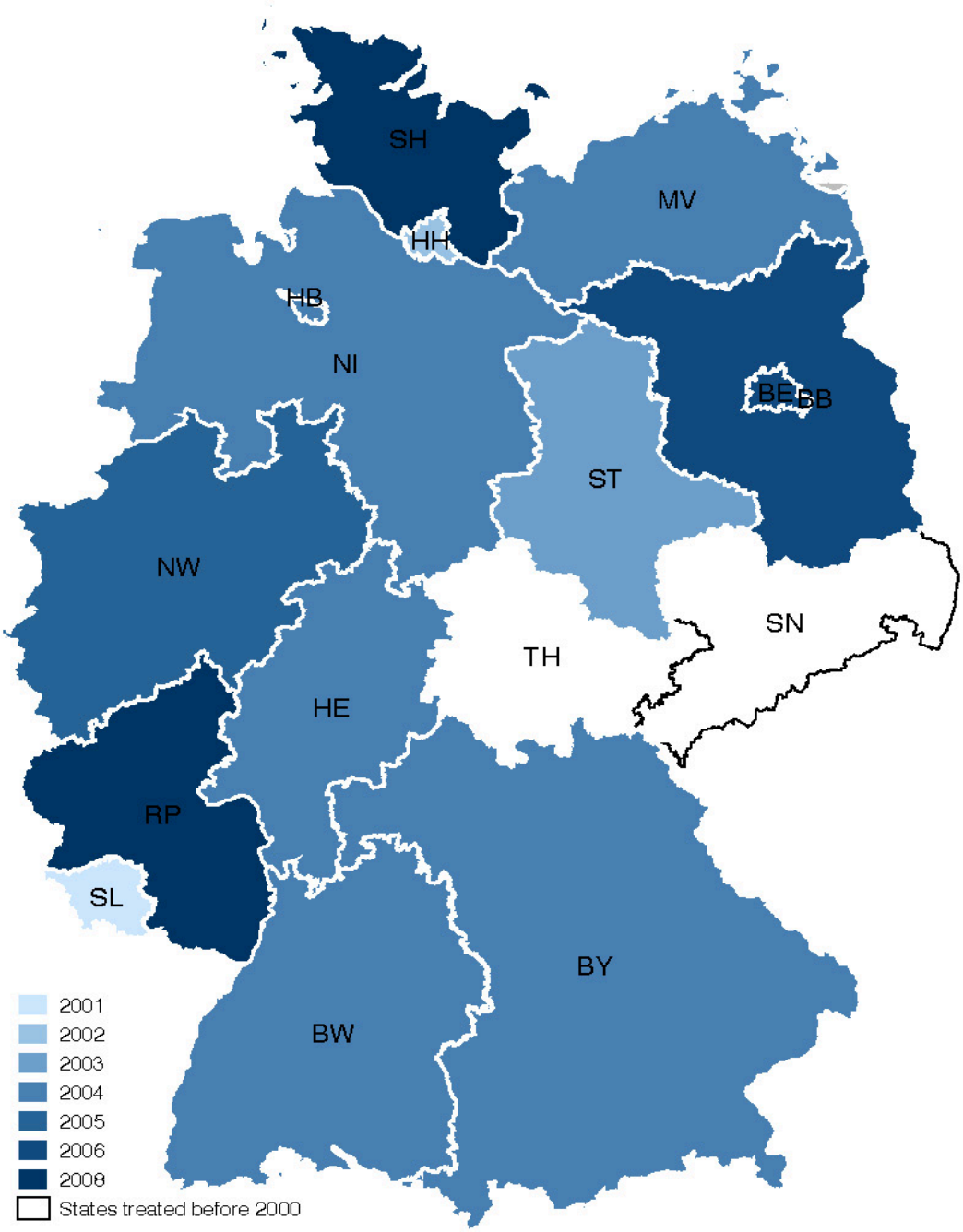
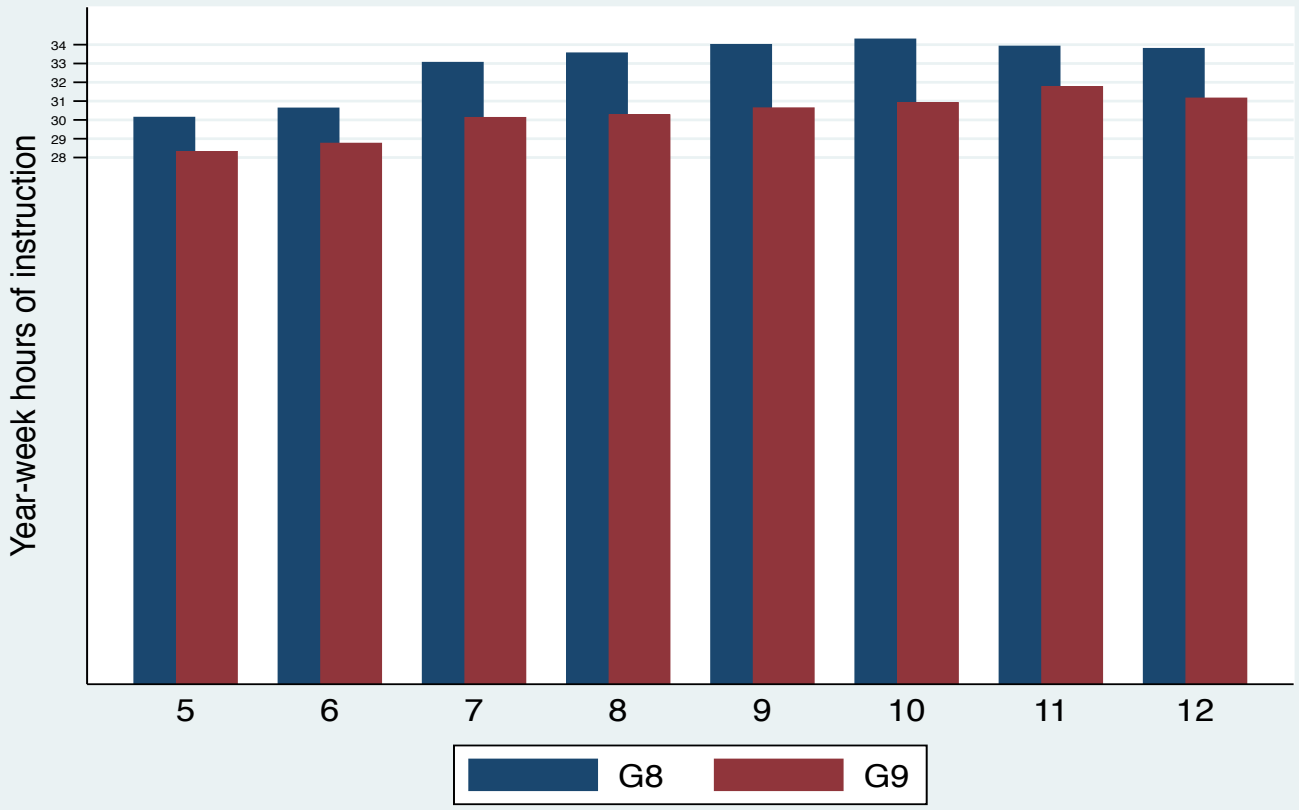


Fig. 3. G8 vs. G9: average year-week hours of instruction by grade (grades 5-12)



Source: Own elaborations on KMK (1997-2014) data

Table 1. G8 treatment status of PISA cohorts

State	G8 adoption	Grades treated	Tracking year	PISA cohorts 2000	2003	2006	2009	2012
Baden-Württemberg (BW)	2004	**5**	**2004**	C	C	C	**T**	T
Bavaria (BY)	2004	6	2003					
		5	**2004**	C	C	C	T	T
Berlin (BE)	2006	**7**	**2006**	C	C	C	**T**	T
Brandenburg (BB)	2006	**7**	**2006**	C	C	C	**T**	T
Bremen (HB)	2004	**5**	**2004**	C	C	C	**T**	T
Hamburg (HH)	2002	**5**	**2002**	C	C	**C**	T	T
Hesse (HE)*	2004	**5**	**2004**	C	C	C	**C**	T
Mecklenburg-Vorpommern (MV)**	2004	9	2000					
		8	**2001**	C	C	T**	T	T
		7	2002					
		6	2003					
		5	**2004**					
Lower Saxony (NI)	2004	6	2003					
		5	**2004**	C	C	C	T	T
North Rhine-Westfalia (NW)	2005	**5**	**2005**	C	C	C	**C**	T
Rhineland-Palatinate (RP)***	2008	**5**	**2008**	C	C	C	C	**C**
Saarland (SL)	2001	**5**	**2001**	C	C	**T**	T	T
Saxony (SN)****	1992	5	1992	T	T	T	T	T
Saxony-Anhalt (ST)**	2003	9	1999					
		8	2000					
		7	**2001**	C	C	T**	T	T
		6	2002					
		5	2003					
Schleswig-Holstein (SH)	2007	**5**	**2007**	C	C	C	C	**C**
Thuringia (TH)****	1991	5	1991	T	T	T	T	T

Notes: Column 1 indicates the year when the G8 reform was adopted. Column 2 reports the grades (cohorts) initially treated. Column 3 reports the tracking year of the cohorts initially treated, i.e., the academic year in which they entered academic-track high school. Figures in columns 2 and 3 are reported in bold when relevant to define the treatment status of PISA cohorts. T and C indicate treatment and control group, respectively; when they are reported in bold, they indicate that the cohort observed is the first G8 cohort or the last G9 cohort, respectively. * In Hesse the G8 reform was introduced gradually: 10%, 60%, and 30% of schools were affected in 2004, 2005, and 2006, respectively. ** The PISA 2006 cohorts In Mecklenburg-Vorpommern and Saxony-Anhalt entered academic-track high school in 2001 (see Figure 4), and were therefore treated only in grades 8 to 9 and 7 to 9, respectively. *** In Rhineland-Palatinate the reform has only been introduced in selected schools so far. **** After reunification, Saxony and Thuringia kept the G8 regime that was typical of academic-track high schools in former East states. *Source*: Kulturministerkonferenz (KMK).

Fig. 4. Academic-track high school enrollment by PISA cohort

	Grade attended by year						PISA 2000
Grade 8			5	6	7	8	
Grade 9		**5**	**6**	**7**	**8**	**9**	**PISA c.**
Grade 10	5	6	7	8	9	10	**test**
Year	1994	1995	1996	1997	1998	1999	2000

	Grade attended by year						PISA 2003
Grade 8				5	6	7	8	
Grade 9			**5**	**6**	**7**	**8**	**9**	**PISA c.**
Grade 10	5	6	7	8	9	10	**test**
Year	1997	1998	1999	2000	2001	2002	2003

	Grade attended by year						PISA 2006
Grade 8				5	6	7	8	
Grade 9			**5**	**6**	**7**	**8**	**9**	**PISA c.**
Grade 10	5	6	7	8	9	10	**test**
Year	2000	2001	2002	2003	2004	2005	2006

	Grade attended by year						PISA 2009
Grade 8				5	6	7	8	
Grade 9			**5**	**6**	**7**	**8**	**9**	**PISA c.**
Grade 10	5	6	7	8	9	10	**test**
Year	2003	2004	2005	2006	2007	2008	2009

	Grade attended by year						PISA 2012
Grade 8				5	6	7	8	
Grade 9			**5**	**6**	**7**	**8**	**9**	**PISA c.**
Grade 10	5	6	7	8	9	10	**test**
Year	2006	2007	2008	2009	2010	2011	2012

Table 2. Summary statistics

Variable	Mean	SD
PISA scores		
Reading	572.13	55.51
Mathematics	578.08	58.61
Science	586.10	61.70
Student controls:		
Demographics		
Female	0.53	0.50
Age (in months)	185.22	5.54
Grade repeated	0.08	0.27
Socio-economic background		
Parents' ISCED 3-4	0.29	0.45
Parents' ISCED 5-6	0.62	0.49
Parents' ISEI	59.25	17.34
Books in house: >100	0.58	0.49
Only child	0.29	0.45
Kid born in foreign country	0.04	0.20
Parents born in foreign country	0.13	0.34
No German spoken at home	0.04	0.20
School controls:		
School enrollment	793.93	352.15
% of girls enrolled	49.42	15.07
Student-teacher ratio	14.66	5.93
Lack of computers	0.33	0.47
Lack of textbooks	0.23	0.42
Urban school	0.26	0.44
Private school	0.08	0.26
Policy variables		
G8 reform	0.41	0.49
Duration of treatment	1.61	2.30
Avg. year-week hours of instruction (grades 5-9)	30.93	1.49
Observations	33,996	

Notes: The sample includes academic-track ninth-graders from PISA 2000-2012 pooled data with a valid assessment in reading. Final student weights are used. Descriptive statistics reported for mathematics scores are based on the mathematics sample (N=29,929). Descriptive statistics reported for science scores are based on the science sample (N=30,202).

Table 3. Main results

	(1)	(2)	(3)
Panel A: Reading			
G8 reform	0.073**	0.079**	0.073**
	(0.022)	(0.021)	(0.024)
Duration of treatment	0.013*	0.014**	0.013*
	(0.005)	(0.005)	(0.006)
Avg. year-week hours of instruction (grades 5-9)	0.031**	0.034***	0.032***
	(0.010)	(0.008)	(0.009)
Observations		33,996	
Panel B: Math			
G8 reform	0.075*	0.079**	0.073**
	(0.044)	(0.035)	(0.035)
Duration of treatment	0.015	0.016*	0.014*
	(0.010)	(0.008)	(0.008)
Avg. year-week hours of instruction (grades 5-9)	0.023*	0.024**	0.023**
	(0.014)	(0.010)	(0.010)
Observations		29,929	
Panel C: Science			
G8 reform	0.088**	0.088**	0.087**
	(0.026)	(0.019)	(0.020)
Duration of treatment	0.017**	0.016**	0.016**
	(0.006)	(0.005)	(0.005)
Avg. year-week hours of instruction (grades 5-9)	0.026**	0.027**	0.027**
	(0.011)	(0.009)	(0.009)
Observations		30,202	
Cohort fixed effects	✓	✓	✓
State fixed effects	✓	✓	✓
Student controls		✓	✓
School controls			✓

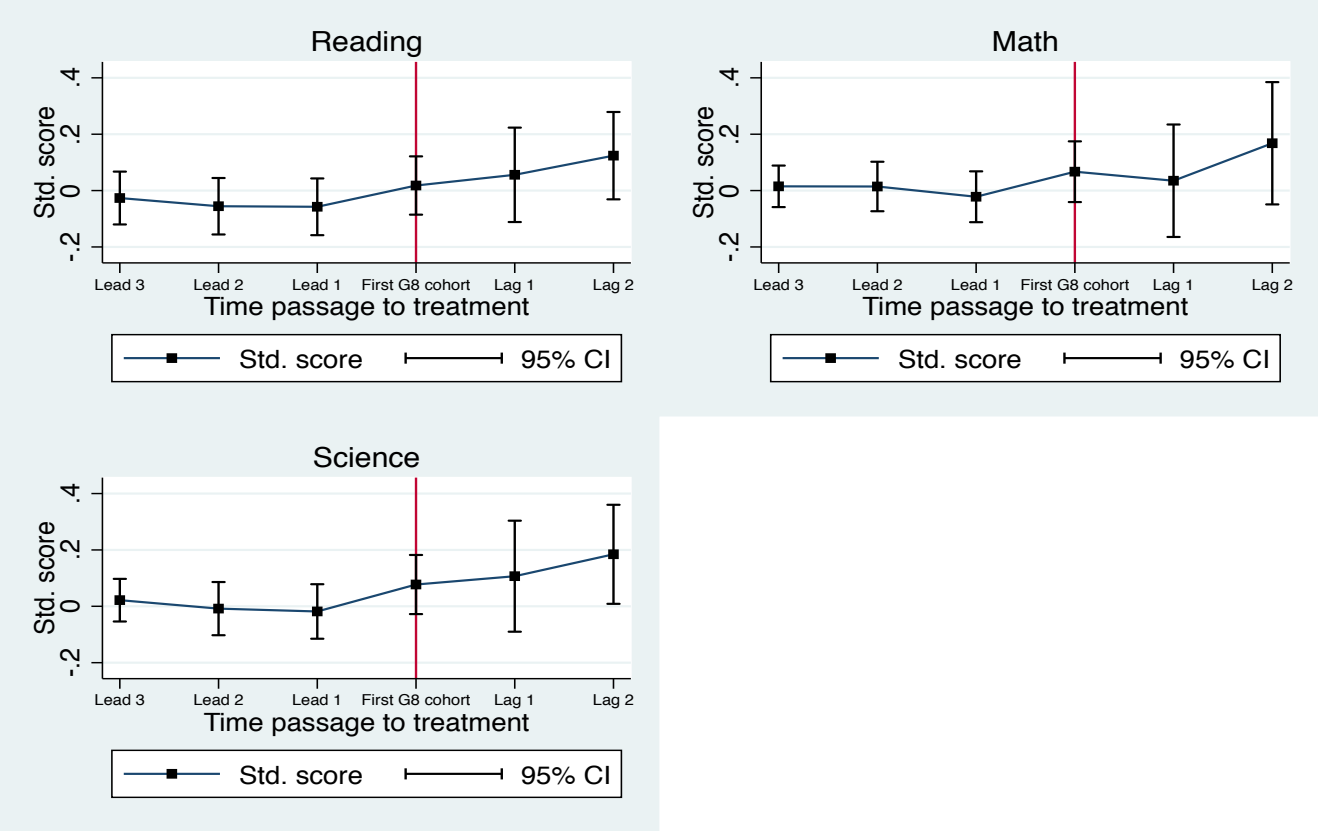
Notes: OLS coefficients (and standard errors) on G8 reform, duration of treatment, and average year-week hours of instruction allotted to grades 5-9 reported, respectively, in the first, second, and third row of each panel are estimated separately from equation (1). Specification (1) is the baseline specification. Specification (3) is the main specification, including student and school controls reported in Table 2. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The samples in panel A, B, and C include academic-track ninth-graders from the pooled PISA 2000-2012 dataset with a valid assessment in either reading, math, or science, respectively.

Table 4. Heterogeneous effects

	Female gender (1)	High educ parents (2)	Migrant parents (3)	Grade retention (4)
Panel A: Reading				
G8 reform	0.000	0.102**	0.079**	0.083**
	(0.026)	(0.046)	(0.021)	(0.021)
Interaction	0.137**	-0.045	-0.063	-0.133**
	(0.022)	(0.043)	(0.051)	(0.059)
Observations		33,996		
Panel B: Math				
G8 reform	0.086*	0.091*	0.078**	0.081**
	(0.047)	(0.049)	(0.034)	(0.034)
Interaction	-0.029	-0.029	-0.067*	-0.121**
	(0.038)	(0.038)	(0.039)	(0.040)
Observations		29,929		
Panel C: Science				
G8 reform	0.069**	0.108**	0.099**	0.099**
	(0.025)	(0.037)	(0.019)	(0.020)
Interaction	0.034*	-0.032	-0.138**	-0.166**
	(0.018)	(0.040)	(0.045)	(0.033)
Observations		30,202		

Notes: All estimated models based on the main specification – i.e., specification (3) in Table 3 – and include an interaction term between the column category dummy and the G8 dummy. High educated parents have ISCED equals to five or six. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. Panel A, B, and C samples include ninth-graders in academic-track high schools from the pooled PISA 2000-2012 dataset with a valid assessment in either reading, math, or science, respectively.

Fig. 5. Standardized test scores and the G8 reform



Source: Computations on PISA 2000-2012 pooled data (baseline specification, final student weights used)

Table 5. Robustness analysis

	Main spec. (1)	DD lead (2)	Placebos lower-tracks (3)	State trends (4)	DDD model (5)	DDD placebo (6)	Academic track (7)	Switch to CEE (8)	Double cohorts (9)	Without BE, BB, HB, NI, MV, ST (10)	Without SN, TH (11)
Panel A: Reading											
G8 reform	0.073**	-0.012	0.008	0.071**	0.090	0.031	0.012	0.067**	0.077**	0.063*	0.081**
	(0.024)	(0.026)	(0.035)	(0.025)	(0.056)	(0.042)	(0.027)	(0.030)	(0.028)	(0.035)	(0.023)
Observations	33,996	33,996	57,748	33,996	72,053	72,053	109,191	33,996	33,996	23,655	30,142
Panel B: Math											
G8 reform	0.073**	-0.036	0.005	0.096**	0.082	-0.001		0.069	0.072**	0.057	0.072**
	(0.035)	(0.043)	(0.030)	(0.032)	(0.053)	(0.041)		(0.048)	(0.036)	(0.048)	(0.032)
Observations	29,929	29,929	50,542	29,929	63,289	63,289		29,929	29,929	20,956	26,647
Panel C: Science											
G8 reform	0.087**	-0.023	0.024	0.098**	0.082**	-0.000		0.091**	0.092**	0.075**	0.083**
	(0.020)	(0.036)	(0.044)	(0.030)	(0.037)	(0.039)		(0.034)	(0.024)	(0.027)	(0.018)
Observations	30,202	30,202	50,988	30,202	63,886	63,886		30,202	30,202	21,165	26,884

Notes: All estimated models based on the main specification – i.e., specification (3) in Table 3. Attendance of academic-track high school (vs. other types of secondary schools) is used as dependent variable in column (7). Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The main samples – used in columns (1), (2), (4), (8), and (9) – include academic-track ninth-graders with a valid assessment in either reading, math, or science (panel A, B, and C, respectively) from the pooled PISA 2000-2012 dataset. The samples used in column (3) include ninth-graders in basic- and middle-track schools from the pooled 2000-2012 PISA dataset. The samples used in columns (5)-(6) include ninth-graders in middle- and academic-track schools from the pooled PISA 2000-2012 dataset. The sample used in column (7) includes the full sample of ninth-graders from the pooled PISA 2000-2012 dataset. The samples used in column (10) exclude the states of Berlin (BB), Brandeburg (BB), Bremen (HB), Lower Saxony (NI), Mecklenburg Vorpommern (MV), and Saxony-Anhalt (ST) from the main samples. The samples used in column (11) exclude the states of Saxony (SN) and Thuringia (TH) (always treated) from the main samples.

Table 6. Linear probability models of high school grade retention and remedial education

	High school grade retention				Remedial education		
		Heterogeneous effects					
	Main spec. (1)	**Female gender (2)**	**High educ. parent (3)**	**Migrant parents (4)**	**Subject: German (5)**	**Subject: Math (6)**	**Subject: Any (7)**
G8 reform	0.001	0.005	-0.005	-0.002	-0.012	-0.079**	-0.067**
	(0.016)	(0.018)	(0.016)	(0.016)	(0.010)	(0.024)	(0.020)
Interaction		-0.006	0.009*	0.050**			
		(0.009)	(0.005)	(0.016)			
Observations			30,490		14,488	9,307	17,655

Notes: Dependent variable in columns (1) - (4) equals one if a grade was repeated in high school, zero otherwise. Dependent variable in columns (5), (6), and (7) equals one if the student is taking a remedial class in German, math, or in any subject, respectively, zero otherwise. Results reported in column (1), and in columns (5) - (7) are based on the main specification, i.e., specification (3) in Table 3. Results reported in column (2) - (4) are obtained estimating models that add to the main specification an interaction term between the column category dummy and the G8 dummy. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The sample used in columns (1) - (4) includes academic-track ninth-graders with non missing values on the dependent variable, and with a valid assessment in reading. The samples used in columns (5), (6), and (7) include academic-track ninth-graders with non missing values on the dependent variable from PISA 2000, 2009, and 2012 cohorts, PISA 2003 and 2009 cohorts, and PISA 2000, 2003, and 2009 cohorts respectively.

Appendix

Table A.1. G8 reform effects on year-week hours of instruction

	Avg. Grades 5-9 (1)	Grade 5 (2)	Grade 6 (3)	Grade 7 (4)	Grade 8 (5)	Grade 9 (6)
G8 reform	2.473**	1.737**	1.235**	3.006**	3.137**	3.253**
	(0.163)	(0.361)	(0.317)	(0.435)	(0.346)	(0.360)
Observations			33,996			

Notes: Dependent variable in columns (1): avg. year-week hours of instruction allocated across grades 5-9. Dependent variables in columns (2) to (6): grade-specific year-week hours of instruction (grade 5 to grade 9, respectively). OLS baseline regressions estimated on the sample of academic-track ninth-graders from the pooled PISA 2000-2012 dataset with a valid assessment in reading. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively.

Table A.2. Standardized annual gains in PISA test scores

	Reading	Mathematics	Science
Grade 8-9	0.88	0.86	0.81
Observations	61,353	49,522	49,541
Grade 9-10	0.50	0.56	0.47
Observations	66,484	53,489	53,553

Notes: Standardized average annual gains in reading, mathematics, and science computed on the PISA 2000-2003 pooled sample of 15-year-old eight, ninth, and tenth graders.

Table A.3. Student controls as dependent variables

	Gender (female) (1)	Age (months) (2)	Parents' ISCED 5-6 (3)	Parents' ISEI (4)	Books in house (>100) (5)	Only child (6)	Migration background (7)
G8 reform	-0.009	0.203	-0.022	0.078	0.008	0.026	-0.033
	(0.022)	(0.389)	(0.019)	(0.988)	(0.026)	(0.018)	(0.030)
Observations	33,922	33,996	33,100	33,680	32,774	32,979	33,421

Notes: Dependent variables in columns (1) to (7): gender (female) (1), age (in months) (2), parents' ISCED (5-6) (3), parents' ISEI (4), books in house (>100) (5), only child (6), migration background (7), respectively. All estimated models based on the baseline specification – i.e., specification (1) in Table 3. Final student weights are used in all regressions. ** and * indicate significance at 5 and 10 percent levels, respectively. The samples includes academic-track ninth-graders with a valid assessment in reading and with non-missing values on the dependent variable from the pooled PISA 2000-2012 dataset.

Table A.4. School controls as dependent variables

	School enrollment (1)	% girls enrolled (2)	Student-teacher ratio (3)	Lack of resources (4)	Urban school (5)	Private school (6)
G8 reform	39.948	-0.617	-1.144	0.002	-0.037	-0.009
	(49.033)	(1.591)	(1.255)	(0.002)	(0.072)	(0.024)
Observations	32,234	32,175	31,319	32,985	32,955	32,956

Notes: Dependent variables in columns (1) to (7): school enrollment (1), percentage of girls enrolled (2), student-teacher ratio (3), lack of computer/textbook resources (4), urban school (5), private school (6), respectively. All estimated models based on the main specification – i.e., specification (1) in Table 3. Final student weights are used in all regressions. Standard errors clustered on state are reported in parentheses. ** and * indicate significance at 5 and 10 percent levels, respectively. The samples includes academic-track ninth-graders with a valid assessment in reading and with non-missing values on the dependent variable from the pooled PISA 2000-2012 dataset.